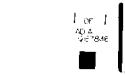
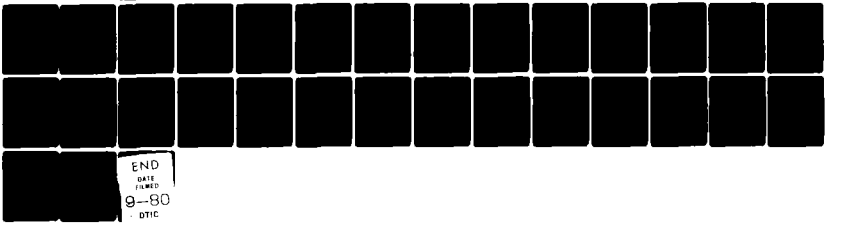
AD-A087 836 CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER F/G 12/1
THE BAYESIAN APPROACH TO STATISTICS.(U)
MAY 80 D V LINDLEY AFOSR-77-3179
UNCLASSIFIED ORC-80-9 NL



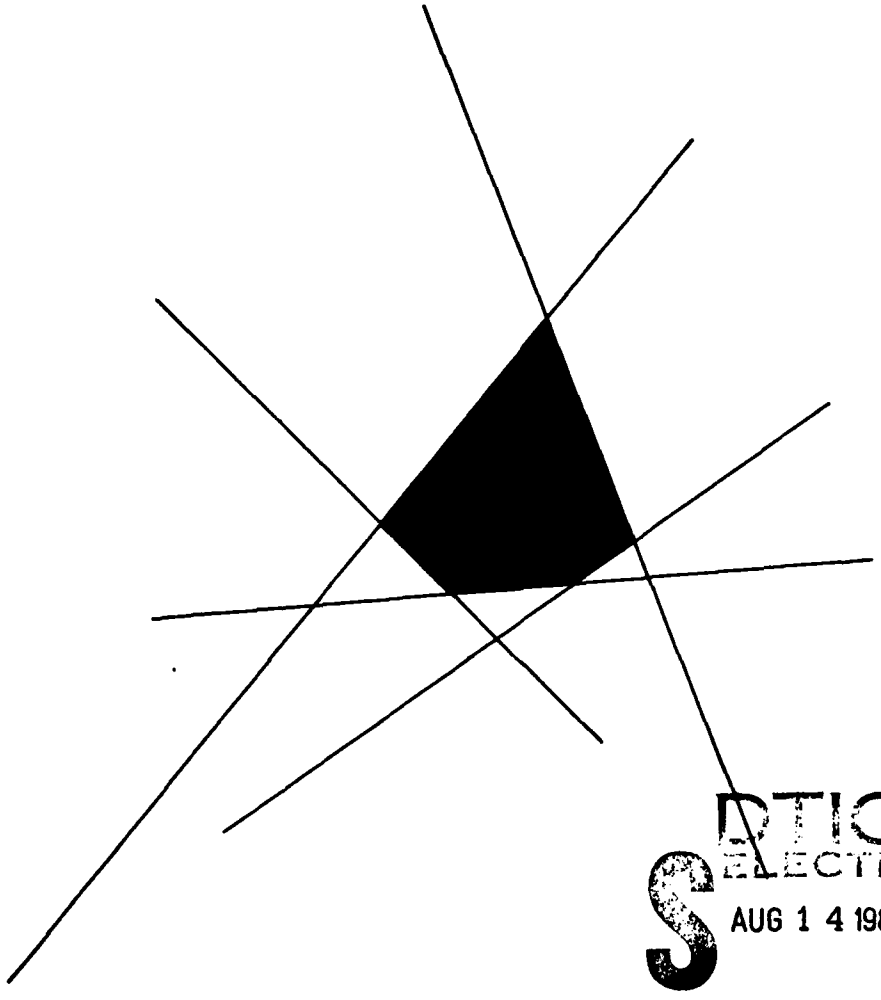


END
DATE FILMED
9-80
DTIC

ORC 80-9
MAY 1980

THE BAYESIAN APPROACH TO STATISTICS

by
DENNIS V. LINDLEY



OPERATIONS RESEARCH CENTER

UNIVERSITY OF CALIFORNIA · BERKELEY

THE BAYESIAN APPROACH TO STATISTICS[†]

by

Dennis V. Lindley[††]

MAY 1980 ORC 80-9

[†]Paper presented at a symposium on recent advances in statistics at Academia das Ciências de Lisboa, April 1980.

This research was supported by the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-77-3179 while the author was a Visiting Professor at the Operations Research Center, University of California, Berkeley. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[††]Present address: _ Periton Lane, Minehead, Somerset TA24 8AQ, England.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ORC 80-9	2. GOVT ACCESSION NO. AD-A087	3. RECIPIENT'S CATALOG NUMBER 836
4. TITLE (and Subtitle) THE BAYESIAN APPROACH TO STATISTICS		5. TYPE OF REPORT & PERIOD COVERED Research Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dennis V. Lindley		8. CONTRACT OR GRANT NUMBER(s) AFOSR-77-3179
9. PERFORMING ORGANIZATION NAME AND ADDRESS Operations Research Center University of California Berkeley, California 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2304/A5 A37
11. CONTROLLING OFFICE NAME AND ADDRESS United States Air Force Air Force Office of Scientific Research Bolling AFB, D.C. 20332		12. REPORT DATE May 1980
		13. NUMBER OF PAGES 41
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Random Quantity; Rules of Probability; Coherence; Bayesian View; Conditional Probability; Borel's Paradox; Scoring Rules; Multiple Comparisons; Independence; σ-fields; Likelihood; Likelihood Principle; Called-off Bets, Models; Parameters; Exchangeability; Hyperparameters; Goodness-of-Fit; Fit of Polynomials; Alternative Hypotheses; Cromwell's Rule; Calibration; Subjectivity; Probability Appraisal.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE ABSTRACT)

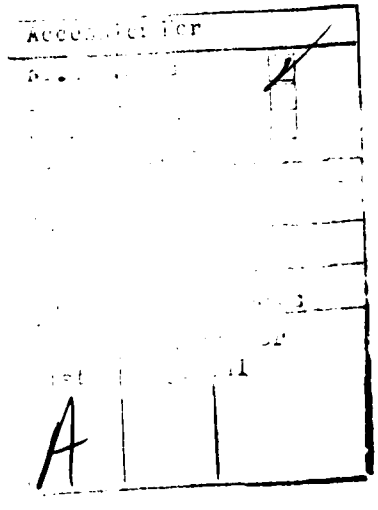
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

This paper discusses several topics that arise in applying Bayesian ideas to inference problems. The Bayesian paradigm is first described as an appreciation of the world through probability: probability being expressed in terms of gambles. Various justifications for this view are outlined. The role of models in the specification of probabilities is considered; together with related problems of the size and complexity of the model, robustness and goodness of fit. Some attempt is made to clarify the concept of conditioning in probability statements. The role of the second argument in a probability function is emphasized again in discussion of the likelihood principle. The relationshiip between the probability specification and real-world experiences is explored and a suggestion is made that zero probabilities are, in a sense, unreasonable. It is pointed out that it is unrealistic to think of probablity as necessarily being defined over a σ-field. The paper concludes with some remarks on two common objections to the Bayesian view.



THE BAYESIAN APPROACH TO STATISTICS

by

Dennis V. Lindley

0. INTRODUCTION

The Bayesian approach to statistics is a complete, logical framework for the discussion and solution of problems of inference and of non-competitive decision-making. It has many facets, from axiomatic foundations of probability to sophisticated technical manipulations needed to solve practical problems. The present paper deals only with inference and is devoted to some questions of Bayesian philosophy. We explore some of the points that arise when the mathematics is related to the real world, avoiding both the completely logical questions within the mathematics and the technicalities of real-world situations, and concentrating on our ways of thinking within the paradigm. A wider view is given in Lindley (1971) with some additional comments in Lindley (1978).

In Section 1 the Bayesian paradigm is defined: essentially as a probabilistic view of the scientific world. The meaning of probability is explained. Practical difficulties in so appreciating the whole of our environment are formidable and we naturally need to view parts in isolation: how this can be done is considered in Section 2. The important role of conditioning, especially on new information, is the topic of Section 3. All statistics, and much of science, uses models to describe phenomena. The meaning and role of models in Bayesian statistics is discussed in Section 4, and continued in Section 5 where the problem of the fit of a model is considered. In Section 6 the relationship of a Bayesian view of the world to the reality of that world is investigated and a suggested strengthening of the paradigm to exclude certain

undesirable behaviour is introduced. In Section 7 replies are offered to two of the most oft-repeated criticisms of the Bayesian attitude.

There is no pretense to completeness in the range of topics discussed: they are points that seem to me to be of some importance and on which I may have a little that is new to say. My debt to de Finetti is considerable. Often, when I feel I have something new to say, I realize that all I am really doing is appreciating the significance of some of his writings. Sometimes, unfortunately, I may have misinterpreted some of his views. The potential reader of this paper might better spend his time with de Finetti (1972, 1974/5).

1. THE BAYESIAN PARADIGM

The scientist's appreciation of the world--and I use the possessive to remind us that poets, musicians, artists and others see the world differently--the scientist's appreciation is of a collection of quantities, of things that can be described by numbers. For a scientist to understand and manipulate things, he must measure them, or at least think of them as things that, indirectly or directly, might be measurable. For example, human preferences are studied by measuring them in terms of quantities called utilities. Without this quantification the scientist cannot proceed: with it, he has at his disposal the full force of logic and the mathematical argument. The success of the scientific approach depends, in part, on how well this quantification encapsulates the situation under study. At the moment it does rather well in mechanics; less well in psychological studies of preference.

These quantities are of two types: those whose numerical value is known to the scientist, and those which are unknown. The members of the first group are familiar to us as known, real numbers. The unknown quantities are more mysterious. We shall refer to them as random quantities (r.q.), though the term, uncertain quantity, is often used. A random quantity will be denoted by a capital letter, say X ; its numerical value by the corresponding lower-case letter, x . Collections of random quantities will not be distinguished notationally from single quantities. When X becomes known then x will be replaced by the revealed number. Thus X may be the breadth of the desk on which I write: x its numerical value in feet. When the desk is measured it may be found that $x = 4$ and X , as a random quantity, becomes known to be 4 feet. An important sub-class of random quantities refer to events, where $X = 1$ if the event is true, and $X = 0$ if it is false.

At any point in time the scientist contemplates a set of quantities some of which are known, some unknown. The values of those in the first set describe part of what he knows at the moment. He will also know about some logical relations that exist between the quantities, both known and unknown: for example, the area of the desk is the product of its length and breadth. This knowledge of logic will not appear explicitly in our notation but its presence must not be forgotten; a point we will return to in Section 2. The set of known quantities will be denoted by H and their values by h. Similarly the set of random quantities will be written, X. Whilst it is easy to describe H, namely as h, more elaborate description is needed for random quantities. Although a random quantity is, by its nature, unknown, it is never completely unknown, in the sense that the scientist knows nothing about it. In the example of X, the breadth of the desk, in feet, one would think 2 or 3 much more reasonable values than 500, or 0.002; and -5 is illogical. It is possible to study random quantities such as X, the desk's breadth, almost as though they were merely letters that the mathematician can so powerfully and profitably manipulate, and forget that X is the representation of a real thing. This forgetfulness leads to difficulties in describing the uncertainty which are substantially diminished as soon as the reality behind the letter is remembered. The scientist's problem is to describe this partial knowledge that he has of the random quantities that interest him.

The Bayesian position is that this uncertainty, or partial knowledge, is to be described in terms of probability. Thus the value of 3 feet is more probable for my desk than is that of 500 feet. In general, with X the set of random quantities and H the set of known ones; the Bayesian, scientific description of the world is a probability distribution of X,

given H : this is written $p(X \mid H)$ and read "the probability of X given H". It is most important to recognize that when it is claimed that the description is in terms of probability it is not merely meant that the description is by means of a number lying between 0 and 1. Much more is intended: namely that different uncertainties are related by the rules of probability. We remind the reader of these rules in the case of random events.

Convexity:

$$0 \leq p(X \mid H) \leq 1 \quad \text{and} \quad p(H \mid H) = 1 .$$

Addition:

If X_1 and X_2 are exclusive, meaning the event, X_1 and X_2 both true, is logically impossible, then

$$p(X_1 \text{ or } X_2 \mid H) = p(X_1 \mid H) + p(X_2 \mid H) .$$

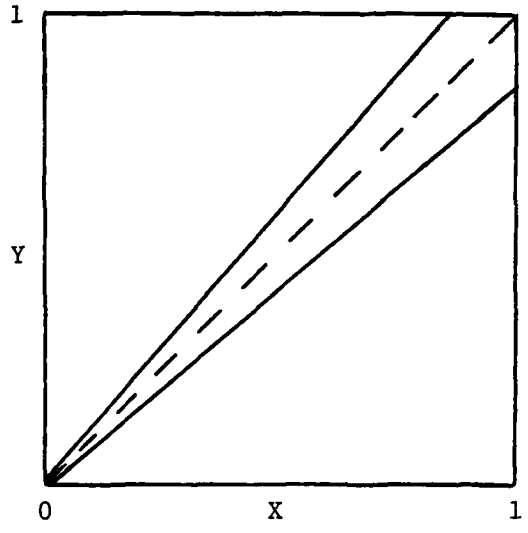
Multiplication:

$$p(X_1 \text{ and } X_2 \mid H) = p(X_1 \mid H)p(X_2 \mid X_1 \text{ and } H) .$$

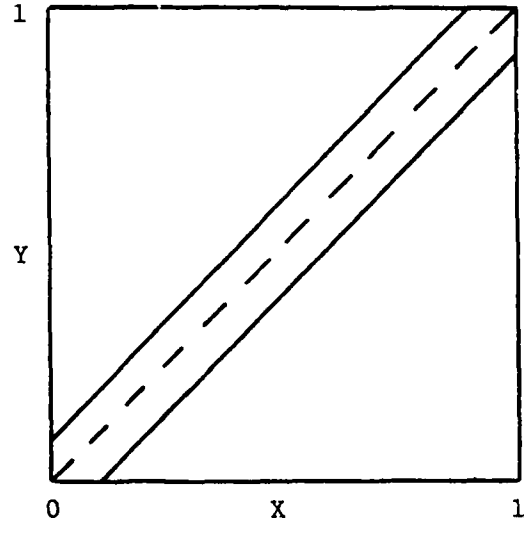
Thus a geneticist who knows that a cross-fertilized plant must be red, pink or white and expresses his uncertainty by saying that the probability of red is 0.4, of pink 0.2, is committed to a probability of 0.6 that the plant will be colored. The addition and multiplication rules describe how uncertainties combine, or cohere, and a set of uncertainty statements obeying the probability rules is often said to be coherent. The Bayesian position is that uncertainties should cohere in this technical sense. Before proceeding further, we digress in the next paragraph to discuss a few points of notation.

In general X and H will be multidimensional and continuous so that $p(X \mid H)$ will be a probability density for X with respect to some dominating measure. To avoid technical problems that are not much the concern of this paper we shall often think in terms of discrete random quantities and use summation in place of the more general integration with respect to the dominating measure. The reader may object that since H is known, as h, it is unnecessary to consider $p(X \mid H)$ and that $p(X \mid h)$ would suffice. There are two reasons for preferring the extended form. Firstly we often need to consider several possible values of H. Thus suppose a scientist is designing a space craft to visit Venus. He is uncertain about the temperature Y on the surface and needs to contemplate the performance X of an instrument in various temperature conditions: thus he contemplates $p(X \mid Y)$ for all temperatures. This point will recur again when we consider in Section 3 more precisely what is meant by "given H" in the probability of X, given H.

The second reason for using H instead of h is technical. If, as in the Venus example, Y is continuous, it transpires that a satisfactory definition of conditional probability is only possible for a random quantity Y and not for a single measurement where the quantity is unspecified. This is usually known as Borel's paradox. To illustrate, let a point, described by its random coordinates X and Y, be uniformly distributed in the unit square. It is meaningless to consider the uncertainty in X given that the point lies on the diagonal through the origin. For if the diagonal is described by $Z = X/Y = 1$ and measurements are of Z, the situation is quite different from when it is described by $W = X - Y = 0$. This can be seen by considering values of Z <u>near</u> 1 and W <u>near</u> 0 : Z near 1 (Figure 1a) means X is more likely to be near 1 than near 0; whereas W near 0 (Figure 1b) means all values are equally likely.



a) Z = X/Y near to 1



b) W = X - Y near to 0

FIGURE 1

A POINT (X,Y) UNIFORMLY DISTRIBUTED WITHIN THE UNIT SQUARE

Still on notation, we shall sometimes write $p_{X,H}(\cdot \mid \cdot)$ instead of the shorter $p(X \mid H)$ to emphasize that we are dealing with functions, in this case of two variables, X and H. For example, $p_{X,H}(2 \mid 3)$, in the case of one random quantity, X and one known, H, means the probability that $x = 2$, given that $h = 3$: here the variables have taken the numerical values 2 and 3.

To return to the Bayesian paradigm, asserting that uncertainties about quantities should cohere according to the rules of probability, we need to ask why they should combine this way. There are two answers, axiomatic and pragmatic. In outline, the axiomatic approach proceeds by searching for simple, self-obvious truths about uncertainty, taking these as axioms and developing a mathematical system of theorems and proofs from them. As an example of such a truth, consider three uncertain events A, B and C for which a scientist thinks that A is more likely than B and B more likely than C. Then it seems self-evident that he should think A more likely than C. From a set of such axioms it is possible to prove that the relationship "more likely than" just mentioned corresponds to an ordering by means of probability, obeying the three rules described above. The first person to attempt such a development successfully was Ramsey (1964)--the original publication was 1931. The first detailed discussion was by Savage (1954). A good, modern exposition is by DeGroot (1970). Many scientific studies use the axiomatic approach, Euclidean geometry and Newtonian mechanics are good examples, because if the axioms encapsulate the real-world situation accurately the theorems will prove useful guides to action. Modern statistical inference outside the Bayesian paradigm lacks such an axiomatic structure.

The above approach is essentially within the field of pure mathematics; in order to use it, to pass to applied mathematics, it is necessary to have an operational meaning for the terms that occur in the mathematics. What do we mean when we say that the probability that $X = 2$, given $h = 3$, is 0.4? How can we operate with such a number? The concept of probability has been much discussed with frequentist, logical and subjective interpretations. Here the subjectivist view will be adopted. Consider an urn containing a white and b black balls and a gamble in which a prize will be won if, on drawing a ball at random from the urn, it is white. Contrast this with a gamble in which the same prize will be won if a random event A is true. If the scientist is indifferent between the two gambles we say the probability for A is $a/(a + b)$. Any increase (decrease) in a will make the urn gamble more (less) attractive. Notice that this interpretation can be tested, to see if the scientist is so indifferent. We do not suggest it is necessarily the best test. Ways of testing, using coherence, have been considered by Lindley et al., (1979).

(Notice that in the interpretation the phrase "drawing a ball at random" has been used. This means that were a prize to be given if a particular ball was drawn, it would not matter which ball was selected to decide the prize. Loosely, all balls are equally likely. The interpretation is not circular.)

The gambling interpretation leads to the pragmatic justification for the Bayesian view: simply that it works. The axioms are prescriptions for one's attitude to uncertainty and they lead to the general, Bayesian prescription or recipe. Let X denote the random quantities of interest: let H denote the known quantities. Then the Bayesian recipe is to calculate $p(X \mid H)$, the calculation to proceed using the

calculus of probabilities: nothing more, nothing less. My claim is that the recipe works: try it for yourselves and see.

An alternative approach which is, in some ways, intermediate between the axiomatic and pragmatic ones, is due to de Finetti (1974/5). Suppose, on considering an uncertain event E, a scientist describes his uncertainty by a number x, knowing that if E subsequently turns out to be true he will receive a penalty score $(x - 1)^2$, whereas if it is false he will receive x^2. Then from the single principle, that if the scientist gives $x,y, \ldots$ for $E,F, \ldots$, the scores being additive, he will not choose values that are such that other values exist for which, whatever is true, the score is reduced, de Finetti shows that the given values must be probabilities. This approach has the merits of the axiomatic approach in being developed from few principles and yet also, in the score, gives an operational interpretation for the value.

According to Popper, a requirement of a good theory is that it should be possible to make predictions which can be tested in practice. If the test succeeds, the theory is enhanced; any failure, and the theory is damaged. Our experience with the probability calculus is extensive and many predictions can be made and tested: there is no case known to me where they fail. This is not to say that the predictions can be evaluated in all cases: on the contrary, there are many situations where the technical difficulties are formidable and unsolved. But this does not invalidate the coherent approach any more than did the initial failure to solve the three-body problem invalidate Newtonian mechanics. In this essay I want to concentrate on the recipe inherent in the Bayesian approach: on the constructive methodology produced by the description of random quantities in terms of probability.

2. LARGE AND SMALL WORLDS

In principle X and H could embrace all that we do not and do know: in practice we confine ourselves to relatively few quantities. It is therefore important to see how contraction can occur. In the case of the random quantities the situation is straightforward. Let X be decomposed into (X_1,X_2) and suppose that only X_1 is of interest. Then by the addition rule

$$p(X_1 \mid H) = \sum_{X_2} p(X_1,X_2 \mid H)$$

to obtain the marginal distribution of X_1. This summation or, in general, integration is basic to the Bayesian recipe and is essentially the addition rule. An important example of its use is in the elimination of nuisance parameters to concentrate on the parameters of interest. Thus in a large agricultural experiment with many varieties and treatments, we can investigate a pair of varieties by integration over treatments and plot constants that appear in the design. Lest this remark appear trivial, let us point out an important practical application. Consider an experiment to compare several treatments and let X_i denote the uncertain yield from treatment i. Then, after experimentation, one has $p(X_1,X_2, \ldots, X_k \mid H)$ for data H, but one may be interested in comparing a treatment that seems to have done rather well with one that has done badly. From the probability just cited one can calculate $p(X_i - X_j \mid H)$ to effect the comparison. Such techniques are not available in sampling-theory statistics and a body of knowledge has been developed, called multiple comparisons, to deal with the problem. In the Bayesian view, no special techniques are needed and the solution is a straightforward use of the addition rule.

If we need to pass from X_1 back to X we can do this by

$$p(X_1,X_2 \mid H) = p(X_1 \mid H)p(X_2 \mid X_1,H)$$

combining the original marginal for X_1 with $p(X_2 \mid X_1,H)$. This is the multiplication rule.

Comparable changes in the known quantities are not so transparent. Let H be decomposed into H_1, H_2. Then

$$p(X \mid H_1) = \sum_{H_2} p(X \mid H_1,H_2)p(H_2 \mid H_1) ,$$

which involves the previously unconsidered $p(H_2 \mid H_1)$, a number that may be hard to assess since it involves forgetting the known quantities, H_2. There is one case where the reduction is simple, namely when

$$p(X \mid H_1,H_2) = p(X \mid H_1)$$

or X is independent of H_2, given H_1. The assumption of independence is often made because it obviates the need to consider $p(H_2 \mid H_1)$, or even H_2, at all. Often it is made without the scientist consciously thinking about it, as when he ignores information in a colleague's paper, H_2, when considering his own. One of the most attractive features of the Bayesian recipe is its ability to put all information together coherently, rather than consider it piecemeal, so that all assumptions of independence need to be considered carefully.

An expansion from H_1 to $H = (H_1,H_2)$ usually arises in studying the most important tool of science, experimentation or observation. The basic idea behind observation is that a previously unknown, or random, quantity becomes known. In our term, it passes from the first

argument of a probability function to the second argument. Let (X_1,X_2) be random quantities and suppose X_2 is observed, then

$$p(X_1 \mid X_2,H) = p(X_1,X_2 \mid H)/\sum_{X_1} p(X_1,X_2 \mid H)$$

provided a rule for the passage from the original uncertainty of both X_1 and X_2 to that of X_1 given X_2. Notice that no new uncertainties are required, only the calculation of the margin for X_2. This result is more often used by applying the product rule to the numerator to give

$$p(X_1 \mid X_2,H) = p(X_2 \mid X_1,H)p(X_1 \mid H)/\sum_{X_1} p(X_1,X_2 \mid H) .$$

This is Bayes rule. It is perhaps the most important rule for human understanding of the real world, for it tells us exactly how we should incorporate the observation of X_2 into our scientific appreciation of that world. Our fuller notation, with X_1 as the variable, expresses it more clearly

$$p_{X_1}(\cdot \mid x_2,H) \propto p_{X_2}(x_2 \mid \cdot,H)p_{X_1}(\cdot \mid H) ,$$

the omitted constant of proportionality not involving the variable. The fundamental role played by Bayes result gives the recipe its name: it might be more sensible to term it the <u>coherent recipe</u> since Bayes was only the first of many who played a role in its development.

We have seen how X and H can be changed. Let us now consider how extensive to make the sets of random and known quantities: how big should we make our scientific world? In an idealization of scientific behaviour, the world could be chosen to embrace everything, known and unknown.

Such a scientist would never need to rethink his understanding of the world, for as soon as X_2 , part of $X = (X_1,X_2)$, becomes known he could calculate $p(X_1 \mid X_2,H)$ by Bayes rule. Such a situation is rarely practicable though it does reasonably arise in one problem of some importance, namely the study of a finite population, where $X = X^{(n)} = (X_1,X_2, \ldots, X_n)$, X_i being the random quantity for the i^{th} member of the population. Observation of a sample from the population changes $p(X^{(n)} \mid H)$ into $p\left(X_{m+1}, \ldots, X_n \mid X^{(m)},H\right)$ where $X^{(m)}$ is the sample.

That situation is unusual in that initial specification of all probabilities is a practical possibility. Usually the scientist prefers to specify such probabilities as seen to him at the time to be relevant and only consider others when the need arises. Let us take a simple case of this and explore it in some detail, since that will reveal other important features of the probabilistic understanding of the world.

Suppose that a scientist considers two events A and B . He attaches probabilities to these, given H , $\alpha = p(A \mid H)$ and $\beta = p(B \mid H)$. The only constraints imposed by the coherence requirements are that both α and β lie in the unit interval. Furthermore there is no obligation on the scientist to consider other events logically derivable from A and B : for example, the union of A and B , $A \cup B$. It is usually stated that the events under consideration must form a σ-field, closed under countable unions and intersections. In a sensible description of reality this is not so, the events may have any structure that the scientist finds convenient; and all that the coherence requirements demand is that those probabilities he has <u>actually</u> assessed (and not those he <u>might</u> have assessed) obey the rules of the calculus of probabilities. Suppose next that the scientist does some logic and discovers that A and B are exclusive,

$A \cap B = \emptyset$. Then H has changed because of this logical consideration, to H' say, and he could now assess $p(A \mid H')$ and $p(B \mid H')$ as α' and β' respectively. There is no exact connection between the original assessments, α and β , and the revised values, α' and β' , obtained after the logical calculation. All that can be said is that, assuming $\alpha \leq \beta$ without loss of generality, $0 \leq \alpha' \leq \alpha$ and $(\beta - \alpha)/(1 - \alpha) \leq \beta' \leq \beta$. (An explanation for these inequalities follows in a few lines.) Additionally, $\alpha' + \beta' \leq 1$.

If, in addition to A and B , the scientist considers the union and assigns probability $\gamma = p(A \cup B \mid H)$, he has then implicitly assessed $p(A \cap B \mid H)$ to be $(\alpha + \beta - \gamma)$. Consequently if he now performs the logic and finds that A and B are exclusive, he can update his original probabilities, rescaling them to add to one. Hence $\alpha' = (\gamma - \beta)/(1 - \alpha - \beta + \gamma)$ and $\beta' = (\gamma - \alpha)/(1 - \alpha - \beta + \gamma)$. (Since $\max(\alpha,\beta) = \beta \leq \gamma \leq \alpha + \beta$, these equalities provide the inequalities stated at the end of the previous paragraph.) So we see that it is possible for logical calculations to change probabilities according to rules of the calculus, as in this paragraph, or to change them arbitrarily subject only to some inequality constraints, as in the previous paragraph. The role of logical considerations in formulating uncertain opinions is therefore a subtle one that perhaps deserves more consideration than its heretofore been given. De Finetti (1974/5) has investigated the condition under which a set of probabilities, such as α , β and γ above, do constitute a complete set in the sense that all other events, logically derivable from those considered, have their probabilities implied by the rules of the calculus.

An example of incomplete logic that is often quoted concerns the decimal expansion of π , the ratio of the circumference to the diameter

of a circle. For most of us the probability that the millionth digit of π is 7 is 0.1, despite the fact that were we to do the logic--equivalent to finding $A \cap B = \emptyset$ above--we would know the value of the millionth digit. This is a clear example of how H can change by purely logical considerations. Bayes theorem principally describes how H changes by empirical considerations. It is a practical advantage of the coherent scheme that it can embrace situations in which the logic is incomplete so that, for example, not all the events in the σ-field need to be prescribed.

3. CONDITIONING

It should be clear that probability is a function of <u>two</u> arguments, X and H. Most treatments of probability are based on probability as a measure. Probability as a function of X is a measure; but as a function of H, it is not. The measure-theory accounts diminish the role of the second argument and reinforce this by utilizing independence conditions so substantially that manipulations using the second argument are scarcely needed. We have already seen how independence can simplify the reduction of H and we shall see later how it can be used with great advantage in a Bayesian treatment of models in Section 4. Nevertheless independence needs most careful consideration. For example, independence is itself a conditional statement: we cannot say A is independent of B, but only given H; for the independence may fail if H changes. An excellent discussion is given by Dawid (1979).

The role of the second argument in probability is most clearly revealed in Bayes theorem,

$$p_{X_1}(\cdot \mid x_2,H) \propto p_{X_2}(x_2 \mid \cdot,H)p_{X_1}(\cdot \mid H) .$$

These two measures for X_1 are connected by the function $p_{X_2}(x_2 \mid \cdot,H)$ of X_1 in the other argument. Whilst $p_{X_2}(\cdot \mid x_1,H)$ is a measure, $p_{X_2}(x_2 \mid \cdot,H)$ in the theorem is not. It is called a <u>likelihood</u> function; in contrast to a probability (or measure) function. Its importance is considerable because the only feature of the random quantity X_2, that has become known to be x_2, used in Bayes' result is the likelihood function of X_1 for that value x_2 (and not for other values of X_2).

This is natural since X_2 has passed from being random to being known and the previously possible, but now impossible, values are irrelevant. We say the likelihood function of X_1, given x_2, is <u>sufficient</u>. Much of modern statistics denies this sufficiency and requires more from X_2 than the likelihood function: and it is this denial that constitutes the main difference between the Bayesian and other paradigms of statistics. Notice that in accord with the earlier discussion of Borel's paradox, it is not enough to know x_2, we must know which r.q. it was that was seen to have the value x_2. X_2 may not be forgotten entirely.

In considering the role of the two arguments in a probability statement attention needs directing towards the word "given" in the phrase "probability of X, given H". A scientist, at a point in time, will know H and consider X. In that case it is sensible to say "given H" as an abbreviation for "given H is known". But we have seen that he might evaluate $p(X_1 \mid X_2,H)$, still at the point in time where X_2 is unknown. In that case "given X_2" does not mean X_2 is known but rather "were X_2 to be known": the subjunctive tense is the relevant one. The gambling interpretation of a probability has been explained above, and a gamble on X_1, were X_2 to be true, is interpreted as a gamble on X_1 that only pays if both X_1 and X_2 are true: indeed, a form of this identification constitutes one of the axioms in the theoretical development of the coherent paradigm. It is sometimes referred to as the axiom of <u>called-off bets</u>, meaning that the gamble is called off if X_2 (as an event) is false. Hence in contemplating $p(X_1 \mid X_2,H)$ the scientist is thinking of his attitude to X_1 <u>were</u> X_2 true. There is a distinction between the contemplation of X_2 and the actual experience of X_2. There is no reason why, when X_2 is actually realized and the

subjunctive becomes unnecessary, the scientist should not express his uncertainty about X_1 differently from the original $p(X_1 \mid X_2,H)$. He must still adhere to the original gamble but his attitude to new gambles might be different. All of us experience the situation in which the experience of something changes our view. Thus I might evaluate my probability of something were it to rain tomorrow, but change it when I experience the rain. This does not invalidate the necessity for the original judgements to cohere, but gives a license to change when contemplation becomes reality. What happens here is that the scientist would say that it is not merely the rain that I have experienced but other, previously unconsidered events, like wet clothes, that I had not contemplated. In other words, that my original collection of random quantities was too small. We must always be alert to a possible useful enlargement of the sets of quantities being considered.

Another point that arises here is the difference between "F is true" and "F is known to be true". Really, E given F means the contemplation of E "were F known to be true". This is clear because the probabilistic description of the world is of an incomplete world with lots of true and false events whose truth or falsity is not known to the scientist, so that the probability must incorporate this lack of knowledge. An interesting example of the need for the distinction arises in legal applications. Suppose that a crime has been committed and the forensic evidence establishes the fact that the criminal has a certain property, A . We later learn that an individual, Smith, has A : what is the probability that Smith is the criminal? We need to calculate the probability that "Smith is guilty" given "we know that Smith has A". As soon as one

expresses it this way, one realizes that it might be pertinent to ask how it is that we know Smith has A. One possibility is that when the forensic evidence was described to a police officer he reported that he knew a man with A, his name was Smith. Another possibility is that the police searched amongst their files, or amongst possible suspects, until they found someone with A, who chanced to be Smith. Detailed calculations show that the two methods of acquiring the knowledge that Smith has A lead to different numerical values for the required probability of Smith's guilt.

The point is related to Borel's paradox where we conditioned not just on x but on a random quantity X that has taken the value x. We cannot condition on Smith's having A but on the observation of some random quantity that takes that value. The quantity in a police search is different from that of a policeman's recollection. It is also related to the question of whether a quantity is controlled or not. Thus an experiment might be performed in which a quantity X varies from unit to unit. Such an experiment may give us valuable information for a new unit in which X is uncertain, but little, or no, information for a new unit controlled, or made, to have a prescribed value for X. To appreciate what happens when control is exercised one has to experiment with control. Free and controlled variation differ not in x, but in the X realized to be x.

4. MODELS

The interpretation of probability in terms of a gamble is only meaningful if the gamble can be settled; that is, if the status of the random quantity can be changed to a known quantity either by observation or by logic. The same remark applies to de Finetti's device of a scoring rule. This is no real restriction since it is idle to talk about things that cannot affect observable quantities. It is idle to discuss the random event that Shakespeare wrote the plays attributed to him unless the event has realizable effects--and doubtless the tourist industry in Stratford feels that it has. We like to talk about unobservables because they can influence observables about which gambles can be settled. A more important reason is that the introduction of unobservables can simplify our probability considerations. Let us see how this can happen.

The specification of $p(X \mid H)$ is difficult if only because the dimensionalities of X and H are both large, and we have seen that there are advantages in considering large sets. We therefore seek ways of simplifying the specification. We can always write

$$p(X \mid H) = \sum_{\theta} p(X \mid \theta, H) p(\theta \mid H)$$

for any quantity θ. Suppose now that X is independent of H, given θ; then our assessment of the uncertainty about X can be changed to one conditional on θ but for which H is irrelevant, and to another about θ. In particular if θ can have low dimensionality, the specifications required may be simpler. We say $p(X \mid \theta)$ is then a (probability) <u>model</u> and refer to θ as a <u>parameter</u> of the model. A model is self-contained in the sense that it is uninfluenced by the known quantities

in H and, more importantly, by a change in H. If X is decomposed into (X_1,X_2) we may write

$$p(X_1,X_2 \mid H) = \sum_{\theta} p(X_1 \mid \theta)p(X_2 \mid \theta,X_1)p(\theta \mid H) .$$

If the further assumption is made that θ known not merely suppresses H but also an addition of H to $H \cup X_1$, we may write

$$p(X_1,X_2 \mid H) = \sum_{\theta} p(X_1 \mid \theta)p(X_2 \mid \theta)p(\theta \mid H) \tag{4.1}$$

resulting in yet further simplification since X_1 and X_2 necessarily have smaller dimensionalities than X. Most of the models used in statistics have this further property that make not only X and H independent given θ, but also X_1 and X_2 independent given θ. Notice that "given" is again to be interpreted here in the subjunctive form since the parameter is rarely an observable quantity that can be made known. Rather its role is to stand between X and H to give them, and the components of X, independence properties. A further simplification that is often made is to suppose in (4.1) that the distributions of X_1 and X_2 given θ are functionally the same: in our fuller notation $p_{X_1}(\cdot \mid \theta) = p_{X_2}(\cdot \mid \theta) = f(\cdot \mid \theta)$ say. Generalizing from 2 to n random quantities we have the common statistical model with

$$p(X^{(n)} \mid H) = \sum_{\theta} \prod_{i=1}^{n} f(X_i \mid \theta)p(\theta \mid H) . \tag{4.2}$$

It is usual to make inferences about the parameter

$$p(\theta \mid X^{(n)},H) \propto \prod_{i=1}^{n} f(X_i \mid \theta)p(\theta \mid H)$$

but it is often more realistic to make inferences in terms of observables. Thus we may write

$$p\left(X_n \mid X^{(n-1)}, H\right) = \sum_{\theta} f(X_n \mid \theta) p(\theta \mid X^{(n-1)}, H) \qquad (4.3)$$

in virtue of the independence of X_n and $X^{(n-1)}$, given θ. The role of the model and of the parameter is to simplify the calculations and to isolate extraneous factors in H from the rest of the system. In particular, the specification of the probabilities given H are reduced from $p(X \mid H)$ to $p(\theta \mid H)$. Even the latter may be too complicated and we may prefer to model that through

$$p(\theta \mid H) = \sum_{\phi} p(\theta \mid \phi) p(\phi \mid H)$$

and <u>hyperparameters</u> ϕ; with θ independent of H, given ϕ. The process may be repeated to model the hyperparameters; and so on. The device is essentially one to simplify our calculations and descriptions.

A sequence $X^{(n)}$ satisfying (4.2) has the property of <u>exchangeability</u> given H: that is, the probability is invariant under permutation of the suffixes $1,2, \ldots, n$, as is easily seen because of the common function f. (Not all exchangeable sequences have the representation (4.2), but the more commonly-used ones do.) This exchangeability implies a particular form of connection between X_n and $X^{(n-1)}$ as spelt out in (4.3). It is an important part of our study of uncertainty to recognize exchangeability and to see, for a random quantity Y, which sequences $X^{(n-1)}$ are exchangeable with Y. This establishes a relation between a random Y and data $x^{(n-1)}$ of known quantities. The point has been considered by Lindley and Novick (1980).

5. MODEL FIT

Although the use of models in expressing uncertainty about quantities is widespread, it has long been recognized that no model is a totally adequate description of one's uncertainty, and that, at best, it is a good approximation. To appreciate this consider the following argument where M is an event of high probability; specifically $p(\bar{M} \mid H) = \varepsilon$ with ε small. Then

$$p(X \mid H) = p(X \mid M,H)p(M \mid H) + p(X \mid \bar{M},H)p(\bar{M} \mid H)$$

or

$$p(X \mid H) - p(X \mid M,H) = \{p(X \mid \bar{M},H) - p(X \mid M,H)\}p(\bar{M} \mid H) .$$

Since the difference of probabilities in braces cannot exceed 2 in modulus,

$$|p(X \mid H) - p(X \mid M,H)| \leq 2\varepsilon .$$

This result enables us to condition on an event, M, of high probability without making much difference to the final result: or we may condition on an event that is nearly true. If M is the event that a model, such as (4.2), is true, then we may evaluate $p(X \mid H)$ as if (4.2) obtained provided we have high probability for the model. If $X = (X_1,X_2)$ and X_2 becomes known, so that H changes to H, X_2, the same argument will persist provided $p(\bar{M} \mid H,X_2)$ remains small. However the observed value x_2 may suggest that M is false and then the calculations will be seriously affected by the supposition of M. Thus the Bayesian paradigm supports the principle of using a simple model until the data suggests it might be false.

One way of proceeding is to think of a model with parameter θ as part of a wider model with parameter (θ,α) that reduces to the earlier model when $\alpha = 0$. If $p(\alpha = 0 \mid H)$ is large we may ignore α, but if x_2 reduces this probability seriously then it may be advisable to consider values of α other than zero. The description of this is not too difficult. One would start, in the usual continuous case, with a density for α centered at zero with small spread and then update this by Bayes theorem to evaluate $p(\alpha \mid X_2,H)$.

It therefore pays to make the model as large as possible. Since an important reason for introducing the model is to simplify the calculations, these two considerations pull in opposite directions and some compromise is essential. There are cases where the technical manipulations are not so hard that very large models may be contemplated. As an example consider a situation where one wishes to make inferences about the relationship between two, one-dimensional quantities Y and X partly expressed through $E(Y \mid X)$. We may model this through $E(Y \mid X,\theta)$ plus other features of the distributions of Y, given X. Over a finite range of X the expectation may be described by an expression

$$E(Y \mid X,\theta) = \sum_{i=0}^{\infty} f_i(X)\theta_i$$

where $f_i(X)$ is a polynomial of degree i orthogonal to the other polynomials and $\theta = \{\theta_1,\theta_2, \ldots\}$. To perform the Bayesian analysis it is necessary to describe the uncertainty about this polynomial expressed through uncertainty about the θ_i's. Our general scientific experience teaches us that polynomials of rather low degrees suffice, so that the distribution for θ_i with large i would concentrate around zero.

The necessary calculations have been performed by Young (1977). There we have a model which, at least as far as the expectation is concerned, could hardly be any larger and no difficulties over its approximations should arise.

Notice that the adequacy of the model can be investigated within the Bayesian paradigm by the calculation of $p(\bar{M} \mid H,X_2)$, or, in parametric form, using $p(\alpha \mid H,X_2)$. In practice, the first suggestion that M is inadequate to explain the data will arise through informal considerations, such as plotting the residuals, but a formal investigation requires the scientist to think about $\bar{M}$; that is, about alternatives to M . We return to this point later in the section.

This combination of informal suggestion of an alternative to the model and its subsequent, more precise analysis has been discussed because it has been suggested, most recently by Box (1980), that the Bayesian paradigm is not adequate to deal with model criticism and that devices outside the coherent system are necessary. One such device is to test the hypothesis that the model is M by calculating $p(X_2 \mid H,M)$ which express the uncertainty, given M , of the random quantity subsequently observed and using this to obtain $p(E(x_2) \mid H,M)$ where $E(x_2)$ is a set of values of X_2 "more extreme than" the observed x_2 . (A common example of "more extreme than" is the tail of the distribution of a univariate X_2 beyond x_2 .) There is a conceptual connection between the choice of what constitutes more extreme values and the alternative models described by α . Whilst the latter consideration is coherent the former is not. One can easily see this by noting that in order to evaluate $p(E(x_2) \mid H,M)$ more information about X_2 has to be provided beyond x_2 as a known value of X_2 , which is all that the coherent

paradigm requires. In particular, values that could fall in $E(x_2)$ have to be contemplated. As Jeffrey's remarks, what might have happened [to X_2] but didn't hardly seem relevant once $X_2 = x_2$ has been observed. The sufficiency of the likelihood function is being denied.

These considerations are also relevant to the closely related question of robustness of an inference procedure wherein we ask how far the final uncertainty $p(X_1 \mid X_2,H)$ is affected by the choice of model. Again, as far as technical considerations allow, the coherent approach should consider large models where the robustness question looms less large. Notice that the method is quite specific about how the data should be analyzed with any model, large or small; namely by calculation of the final probability using only the rules of the probability calculus. Analysis of the normal, linear model within either the standard or the Bayesian paradigm is straightforward. To extend the model and replace normality with t-distributions would cause substantial difficulties within the standard approach, if only because of the nonexistence of sufficient statistics of small dimensionality, whereas anyone with adequate knowledge of the probability calculus and access to computing facilities could perform the arithmetic for any data set.

One difficulty with the above analysis of the adequacy of a model M is that it forces consideration of alternatives $\bar{M}$ to that model. This is not necessary in the approach using $E(x_2)$, except insofar as $E(\cdot)$ might informally be suggested by alternatives. Differently expressed, the coherent paradigm, by requiring a probability distribution over possible models, is essentially a method of <u>comparison</u> of models, not of the adequacy of a single model. (The point applies equally to values of random quantities.) The following example is designed to show why this

difficulty is a real feature of any study of uncertainty, so that its avoidance using $E(\cdot)$, for instance, may be unsound.

A scientist has an observation which is a finite sequence of zeroes and ones. His model for this is a Bernoulli sequence with chance θ . Contrast two cases. In the first he notices that the sequence has a 0 in every even place and 1 in every odd, an unlikely occurrence under his model. He seeks for an alternative model and, although he cannot specify it tightly, he sees one that could explain the data and has reasonable probability; both $p(X_1 \mid \bar{M},H)$ and $p(\bar{M} \mid H)$ for data X_1 and alternative $\bar{M}$ are not near zero. In the second case he notices that the sequence has a 0 in every place whose order is composite and 1 against every prime. In this case no $\bar{M}$ satisfying the condition in the first case is available. Hence we have two pieces of data, X_1 and X_2 , with $p(X_1 \mid M,H) = p(X_2 \mid M,H)$ very small, yet our reactions to them will be different because in one there exist an alternative with reasonable probability, which might be investigated, whereas in the other no such alternative exists and we can only conclude that an unusual observation x_2 has been observed. Actually all observations have something unusual about them: the key question is whether there is a reasonable alternative to explain the observation.

6. CROMWELL'S RULE

Suppose that a random quantity X can only assume one of n values $x_1, x_2, \ldots, x_n$. An event is an example of this, where n is 2. If another random quantity Y is observed the uncertainty about X will be updated by the usual Bayes result

$$p(X \mid Y,H) \propto p(Y \mid X,H)p(X \mid H) .$$

It immediately follows that if $p(x_i \mid H)$ is zero for any i then $p(x_i \mid Y,H)$ is also zero. Since Y is arbitrary, it follows that if $p(x_i \mid H) = 0$, no evidence whatsoever will alter the probability to any value other than zero. Whilst there is nothing in the Bayesian paradigm, as usually discussed, to rule out this possibility, it seems unreasonably rigid to fly in the face of all evidence, even when it supports strongly the possibility that $X = x_i$. We therefore suggest adding an additional requirement, namely that for quantities taking only a finite set of values, no probability be zero. With this addition it is easy to see that if indeed $X = x_i$, then data can eventually accumulate to make this almost certain. (The requirement might be called Cromwell's rule, since he suggested its equivalent when he advised the Church of Scotland to remember that it might be wrong.) In the extension to continuous X, where we are dealing with densities rather than probabilities, the equivalent requirement seems to be that the density nowhere vanishes.

Cromwell's rule is relevant when we consider the relationship between a Bayesian view of the world and the reality of that world that he learns by experience. We have seen that with a complete probabilistic description $p(X \mid H)$, with $X = (X_1, X_2)$ and X_2 observed, the experience is translated into $p(X_1 \mid X_2,H)$, so that he need only update his probabilities

according to the rules and no rethinking, only calculation is needed. This may be unsatisfactory as the following example shows.

Let X consist of a long, perhaps infinite, sequence of ones and zeros and let $p(X \mid H)$ consist in saying that the sequence is a Bernoulli sequence with chance $\frac{1}{2}$ of one at any place. This is a coherent allocation of probabilities to X, assigning probability $(\frac{1}{2})^n$ to any subsequence of length n. Consequently even if a sequence of 100 zeros is observed the probability of a 1 in the 101^{st} place is still $\frac{1}{2}$. This scarcely seems reasonable and a second, coherent scientist watching this behaviour will describe that view as "unscientific" in that it does not incorporate sensible reaction to the data.

The difficulty can be mitigated by judging the sequence to be, conditional on θ, Bernoulli with chance θ and ascribing to θ a density $p(\theta \mid H)$ which nowhere vanishes--unlike the earlier case where $p(\theta \mid H) = 0$ for all $\theta \neq \frac{1}{2}$. This is an example of a model as defined above. Now with any sequence observed, the probability of a 1 in the next place will be around r/n, where r is the number of 1's in the sequence of length n. The second scientist might regard this appreciation of data as "scientific". But now suppose the observed sequence showed 0's and 1's alternating, the probability for a 1 in the next place will be about $\frac{1}{2}$, since $r = \frac{1}{2}n$, whereas a more appealing value might be around 1 if the observed sequence ended in a 0, and 0 if ending in a 1. Hence even this Bayesian reacts unreasonably to some data.

This can be overcome by supposing the sequence to be a first-order Markov chain with parameters (θ,α), θ being as before. Now reaction to the alternating sequence will be reasonable but a sequence exhibiting only triplets 011, say, will be handled unreasonably. This can be

countered by a second-order chain: and so on. Each case is defective, when viewed in the light of the next case, in that it assigns zero densities. For example, the Bernoulli assignment gave zero density to all but one value of α in the Markov chain description. Cromwell's rule would have avoided the difficulties.

A realistic position seems to be that a coherent view must not assign density zero to any possibility. (Or alternatively zero probability to any open set.) That at any stage it may work with a simplified model, such as a Markov chain, assigning some zero densities, having probability near one; but be prepared to abandon it in the light of unexpected data that suggests the model may have lower probability than thought earlier. (This agrees with our discussion of models above.) This recipe seems to lead to reasonable appreciation of data. Cromwell's rule might therefore be added to our axiom system.

The rule is also related to the phenomenon of calibration that has been usefully studied by experimental psychologists: Lichtenstein et al., (1977) provide an excellent survey. A person is said to be calibrated if, after assigning probabilities to a long sequence of independent events, the frequency of events subsequently found to be true amongst all those assigned probability p, is p; and this for all p. Dawid (1980) has shown that every Bayesian is calibrated with probability one. For example, the scientist with a Bernoulli sequence of chance ½ who does not learn from a long string of zeros is nevertheless calibrated. The "catch" in Dawid's result is that the probability referred to is the Bayesian's own probability, so that the adherent of chance ½ has very low probability for the string of zeros. It is not true that one Bayesian will be calibrated with probability one according to a second Bayesian.

A desirable state of affairs would be that every Bayesian would attach probability one to another being calibrated. A requirement for this again seems to be Cromwell's rule so that if one gives probability greater than zero to some possibility, so will the other. Two such Bayesians will ultimately come to agree in the light of suitable data.

Notice that logic, as distinct from experience, can make a probability zero, or one. Thus lengthy, accurate calculations can establish the true value of the millionth digit of π, and Cromwell's rule does not apply.

7. SUBJECTIVITY AND APPRAISAL

Two, related objections to the Bayesian paradigm are often raised. The first is that the procedure is subjective, the second that the probabilities are unknown. These have been much discussed but their frequent repetition suggests some more consideration might be desirable even although there is little new to say.

Consider two scientists contemplating a random quantity X. Then their probabilities for X might reasonably differ. The usual explanation for part, at least, of the difference is that their current states of information differ: that one is evaluating $p(X \mid H_1)$ and the other $p(X \mid H_2)$. Probability is a function of two arguments. It is therefore an important part of the Bayesian paradigm that information should be shared and that both scientists should contemplate X given $H = (H_1, H_2)$. We saw in Section 2 how this might be done coherently for both of them. It is an advantage of the paradigm that it provides the machinery for doing this sharing of knowledge. Even with the same information the two scientists might still have different probabilities $p_i(X \mid H)$, $i = 1,2$. It is possible to argue that no two scientists have, or could have, identical H's and that differences could still be ascribed to difference in parts of H not previously considered. An extension of this view is to argue that, on proper specification of X and H, all people would agree on $p(X \mid H)$, which may be thought of as a rational, or logical, view, and extends ordinary logic. Attempts to calculate $p(X \mid H)$, at least for simple situations, have not been totally successful but neither have they been total failures. Jeffreys (1967) is a protagonist for this view. Box and Tiao (1973) and Zellner (1971) both use the idea in their books. My own view is

the purely subjective view makes more sense. One disadvantage of the logical view is that it discourages serious contemplation of the probabilities--and even of X and H beyond their symbolic meanings--and adopts, perhaps uncritically, a logical probability. Most attempts at describing a logical probability reduce to considering the case where H is, if not empty, at least small, and $p(X \mid H)$ is a probability for X in a state of "ignorance." However we are never ignorant: as long as the words mean something to us, we know something. At best an "ignorance" probability could only serve as a reference point for other, more realistic, ones.

On the subjective view there is no reason why $p_i(X \mid H)$ should agree for two scientists. However, it must be remembered that we are supposing the two scientists are coherent; that is, they have assessed other probabilities that combine together according to the rules of the probability calculus (Section (1)). Much scientific opinion--and especially that based on orthodox statistical principles that deny the likelihood principle--is currently incoherent, and my conjecture is that some disagreement in views results from the scientist's failure to treat his data coherently. Certainly this is true of some significance test arguments and data-analytic techniques that do not assess evidence by consideration of alternative hypotheses (Section (5)).

The message of this section is that two major impediments to agreement can be removed: use of different H's and the failure of coherence. Yet still two scientists might disagree. But now remember that they have expressed their views in an easily understood way, namely by means of numbers (probabilities), and communication between them is easy because they speak a common tongue. Because of this it is possible for

them to see where their principle differences lie and to concentrate attention on these. To do so may be enough to resolve differences. If it is not, then a possibility is to devise an experiment to reduce disagreement. For example, if one scientist attaches high probability to X_1 and another to X_2, the observation of Y, where $p(Y \mid X_1)$ and $p(Y \mid X_2)$ are agreed to differ substantially, essentially resolves the issue. It is necessary to invoke Cromwell's rule, for with one scientist denying with certainty what another scientist credits with being a possibility, is to deny the possibility of critical experimentation.

We see that the subjective approach has the great strength of encouraging cooperation and discussion amongst scientists and of suggesting new experiments. Always the argument comes back to coherence: to the fitting of judgements of uncertainty together. A scientist doing an experiment has been taught by statisticians that there is a unique, proper analysis of the data. When he attempts a Bayesian analysis he sees that this is not so: the analysis will depend on the "prior", on the probability before seeing the data. This will have come from previous evidence, perhaps of earlier experiments that the scientist did. The apparent merit of a unique analysis denies the proper appreciation of the relationship between experiments and the coherence of judgements concerning them. It denies part of learning from experience.

It is often maintained that the likelihood is objective even if the "prior" is not. This is not so: the probability from which it is derived is subjective, like any other probability assessment. There may be more agreement, usually because of the model feature (Section (3)) of independence from H, given θ, but also because of exchangeability being agreed. But were coherence to be used more rigorously, we would have more information

about reasonable likelihoods and have real evidence for normality or other assumptions.

The second objection is that the probabilities are unknown: loosely, what is the "prior"? The Bayesian paradigm purports to describe how a scientist would wish to behave, rather than how he does behave. To achieve this wish he needs to think probabilistically. Once convinced of this, he needs to assess probabilities, to supply actual numbers. Since he has not been doing this he needs to develop expertise and to develop to tools to assist him. The objector is correct in asking his question: he is incorrect if he thinks that its answer is to come as revelation, without substantial studies by the scientist. He must think about the quantity, about ways of assessing it, and must face up to a research project. Scientists who encounter a measurement problem that they think is worth solution, do not say "I cannot do this," rather they launch a research program to discover how to do it. So we should approach the assessment of probabilities.

Some work on the determination of probabilities has been done using scoring-rules and calibration--two topics mentioned above. Useful as both of these devices are, they fail to exploit the basic concept of coherence: they fail to see how one probability impinges on another. Ways of exploiting coherence have been suggested by Lindley et al., (1979). Essentially, the idea is to ask for sets of probabilities. Thus we might try $p(A \mid B)$, $p(A \mid \bar{B})$ and $p(B)$ for events A and B. The coherent assessor is then committed to $p(B \mid A)$, indeed to all uncertainties concerning A and B. He may feel that the calculated $p(B \mid A)$ is unsatisfactory. He can then modify any or all of the original triplet of values to attain coherence and to agree with his total appreciation of the uncertainties surrounding the events.

Neither of the objections dealt with in this section recognize the power of coherence: the ability of that tool to fit opinions concerning uncertainty together in an unobjectionable way. No other procedure besides the Bayesian method can do this: for that method follows from reasonable assumptions of behaviour.

REFERENCES

BOX, GEORGE E. P. (1980), "Sampling and Bayes' inference in scientific modeling and robustness", J. Roy. Statist. Soc. (to appear).

BOX, G. E. P. and TIAO, G. C. (1973), Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, Mass.

DAWID, A. P. (1979), "Conditional independence in statistical theory", J. Roy. Statist. Soc. B, 41, 1-31.

DAWID, A. P. (1980), "The well-calibrated Bayesian", (to appear).

DeGROOT, M. H. (1970), Optimal Statistical Decisions, McGraw-Hill, New York.

DE FINETTI, B. (1972), Probability, Induction and Statistics: The Art of Guessing, Wiley, New York.

DE FINETTI, B. (1974/5), Theory of Probability (2 volumes), Wiley, New York.

JEFFREYS, H. (1967), Theory of Probability, Clarendon Press, Oxford.

LICHTENSTEIN, S., FISCHHOFF, B. and PHILLIPS, L. D. (1977), "Calibration of probabilities: the state of the art", in Decision Making and Change in Human Affairs, H. Jungermann and G. de Zeeuw (editors), Reidel, Dordrecht, 275-324.

LINDLEY, D. V. (1971), Bayesian Statistics: A Review, SIAM, Philadelphia.

LINDLEY, D. V. (1978), "The Bayesian approach", Scand. J. Statist. 5, 1-26.

LINDLEY, D. V. and Novick, M. R. (1980), "The role of exchangeability in inference", Ann. Statist. 8 (to appear).

LINDLEY, D. V., TVERSKY, A. and BROWN, R. V. (1979), "On the reconciliation of probability assessments", J. Roy. Statist. Soc. A, 142, 146-180.

RAMSEY, F. P. (1964), "Truth and probability", in Studies in Subjective Probability, H. E. Kyburg, Jr. and H. E. Smokler (editors), Wiley, New York, 61-92.

SAVAGE, L. J. (1954), The Foundations of Statistics, Wiley, New York.

YOUNG, A. S. (1977), "A Bayesian approach to prediction using polynomials", Biometrika, 64, 309-317.

ZELLNER, A. (1971), An Introduction to Bayesian Inference in Econometrics, Wiley, New York.